Extraction of Nom text regions from stele images using area Voronoi diagram

Thai V. Hoang, Salvatore Tabbone, Ngoc-Yen Pham

Extraction of Nom Text Regions from Stele Images Using Area Voronoi Diagram

Thai V. Hoang[1,2], Salvatore Tabbone[2], Ngoc-Yen Pham[1]
[1]MICA Research Center, Hanoi University of Technology, Hanoi, Vietnam
[2]Université Nancy 2, LORIA, UMR 7503, 54506 Vandoeuvre-lès-Nancy, France
{vanthai.hoang,tabbone}@loria.fr, ngoc-yen.pham@mica.edu.vn

Abstract

Automatic processing of images of steles is a challenging problem due to the variation in their structures and body text characteristics. In this paper, area Voronoi diagram is used to represent the neighborhood of connected components in stele images containing Nom characters. Body text region is then extracted from stele images by the selection of appropriate adjacent Voronoi regions based on the information about the thickness of neighboring connected components. Experimental results show that the proposed method is highly accurate and robust to various types of stele.

1. Introduction

The cultural inheritance of a country is a public property which represents its collective and alive memory and forms a solid base for its future development. Its digitalization is essential for the maintenance and promotion of cultural diversity in this era of globalization. Today, the exploitation of collections of images of old steles has not yet found a satisfactory answer due to their unstructured nature. This raises a problem in the generation of a database, presented in the form of heterogeneous documents, allowing search for information and navigation within these corpora.

A collection of approximately 30.000 stele images containing Nom characters, a derivative of Chinese which was used in Vietnam before the 20th century, has been prepared by EFEO (The French School of Asian Studies) along with their Vietnamese annotation. There is no need to recognize and translate these Nom characters into Vietnamese because their exact interpretation is still under research today. It is needed, however, to process these stele images and represent characters in a database then Nom linguists can do character retrieval in their work. As the initial step of this process, body text regions need to be extracted from stele images.

Locating text regions in a text-containing image is fundamental for its automatic processing. Text-containing regions in stele images are usually encompassed by loosely structured decoration. Figure 1 shows an example of a stele image in which the body text region lies at the center of the image and is enclosed by decorative patterns. This decoration is not only complex in each stele but also varies among different steles. This makes extraction of body text regions extremely difficult.

Figure 1. A stele image

Some methods have been proposed during the past decades to extract regions of document components such as text, figures, tables, and halftones. Fletcher and Kasturi [3] employed Hough transform to the centroids of the rectangles enclosing connected components in order to group together connected components into logical character string. This method is not suitable for Nom characters composed of unconnected strokes. In [4], Jain and Bhattacharjee proposed a texture method using Gabor filter to segment text from document images. This method is time-consuming and inapplicable to stele images as the decorative texture sometimes resembles the texture of body text region.

Recently, some researchers used Voronoi diagram and Delaunay triangulation for document image analysis. For example, Kise et al. [5] and Lu et al. [6] used Voronoi diagram for page segmentation and word grouping in document images respectively using the distance and area ratio between neighboring connected components. Xiao and Yan [9] employed Delaunay triangulation to extract text regions by representing each connected component by its centroid. However, all of these methods work only with Latin characters and are inappropriate to extract Nom text region from stele images. This is because one Nom character may be composed of several connected components whose centroids, areas and inter-distances do not have any regularity.

In this paper, we propose an efficient method based on connected component analysis for the extraction of body text region from stele images. Area Voronoi diagram is employed to represent the neighborhood of connected components. Based on this representation, the process of body text region extraction can be considered as the grouping of appropriated adjacent Voronoi regions. For this purpose, we utilize one characteristic feature of the standard deviation of the thickness of neighboring connected components.

The remainder of this paper is organized as follows. Section 2 briefly gives a basic definition followed by a constructing method of area Voronoi diagram. Section 3 describes the details of body text region extraction using standard deviation of the thickness of neighboring components. Experimental results are given in Section 4, and finally conclusions are drawn in Section 5.

2. Area Voronoi diagram

2.1. Area Voronoi diagram definition

Let $G = \{g_1, \ldots, g_n\}$ be a set of nonoverlapping components in the two dimensional plane, and let $d(p, g_i)$ be the Euclidean distance between a point p and a component g_i defined by:

$$d(p, g_i) = \min_{q \in g_i} d(p, q) \tag{1}$$

Then the Voronoi region $V(g_i)$ and the area Voronoi diagram $V(G)$ are defined by:

$$V(g_i) = \{p | d(p, g_i) \leq d(p, g_j), \forall j \neq i\} \tag{2}$$

$$V(G) = \{V(g_1), \ldots, V(g_n)\} \tag{3}$$

The boundaries of Voronoi regions, which are always curves, are called *Voronoi edges*, and the points where Voronoi edges meet are called *Voronoi points*.

The Voronoi region of each image component corresponds to a portion of the two dimensional plane. It consists of the points from which the distance to the corresponding component is less than or equal to the distance to any other image components.

2.2. Area Voronoi diagram construction

There are two main approaches that are usually used for the construction of area Voronoi diagram:

- The first approach constructs area Voronoi diagram by means of first generating a point Voronoi diagram for a set of sampled points lying on the boundaries of image components and then deleting Voronoi edges generated from points of the same components [1, 5].

- The second approach constructs area Voronoi diagram by labeling the image components then applying morphological operations to expand their boundaries until two expanding labels meet [7, 8].

The first approach is time consuming as it depends on the number of sampled boundary pixels in which a high sampling rate is preferred to obtain an accurate area Voronoi diagram. We utilize here the second approach to construct area Voronoi diagram of stele images using the Euclidean distance metric [8].

Figure 2 demonstrates the steps of constructing area Voronoi diagram for a stele image containing Nom characters. Original binary image I is given in Figure 2(a). Figure 2(b) demonstrates the Euclidean distance map. The area Voronoi diagram V is shown in Figure 2(c). Voronoi regions with their corresponding connected components are given in Figure 2(d).

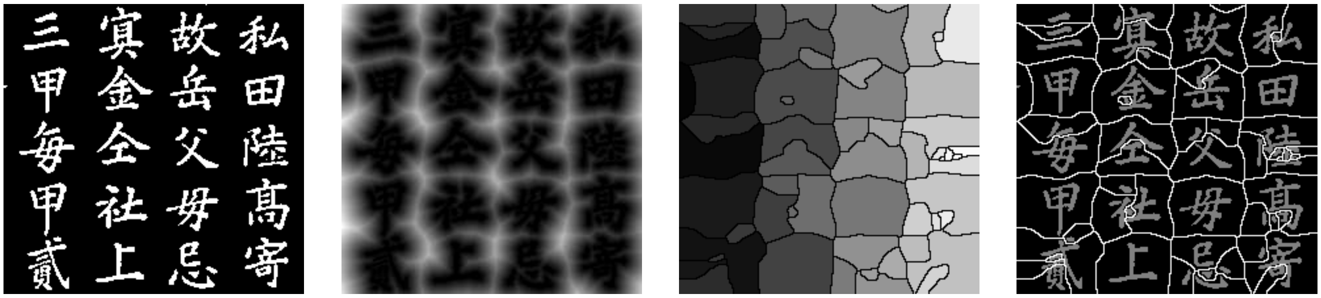
3. Extraction of body text region

As shown in Figure 2(d), each connected component in stele image is represented by one Voronoi region. The process of body text region extraction is then considered as the process of grouping adjacent Voronoi regions which represent the body text region. Therefore, we need criteria for grouping appropriate adjacent Voronoi regions from the area Voronoi diagram.

By observation of Figure 1 we realize that body text region contains connected components of Nom characters that have similar stroke thickness. These components are, in addition, represented by adjacent Voronoi regions. On the contrary, decorative region contains heterogeneous connected components. From this viewpoint, we employ standard deviation of the thickness of neighboring connected components as the discriminating criterion between body text and decorative regions.

3.1. Standard deviation of the thickness of neighboring connected components

In order to calculate the standard deviation of the thickness of neighboring connected components, the thickness



(a) (b) (c) (d)

Figure 2. Construction of area Voronoi diagram

of each connected component need to be calculated first. We propose here a simple but efficient enough method to calculate the thickness of a connected component using Euclidean distance map. The *thickness* of a connected component g_i is approximately defined as two times the average of shortest distances from the pixels lying on its skeleton to its boundary.

In determining the skeletons from the Euclidean distance map obtained in Section 2.2, the algorithm in [2] has been utilized. Figure 3 gives an example of calculating the thickness of a connected component. Original character, its distance map, and its skeleton are given in Figure 3(a), 3(b), 3(c) respectively. The thickness of this connected component has the value 3.61.

(a) (b) (c)

Figure 3. Distance map and skeleton of a single connected component

The standard deviation of the thickness of two neighboring connected components g_i and g_j, by definition, is calculated by:

$$std(g_i, g_j) = \left| \frac{thickness(g_i) - thickness(g_j)}{2} \right| \quad (4)$$

3.2. Grouping adjacent Voronoi regions

According to (4), the value of std among body text region is small while its value among decorative region is highly diverse. Figure 4 gives the histogram of std extracted from the stele image in Figure 1. There are two distinct clusters in the histogram:

- The cluster of higher value of std corresponds to neighboring connected components with high variance in their thickness and therefore these components belong to the decorative region.
- The cluster of lower value of std corresponds to neighboring connected components with small variance in their thickness, however these components may belong to the body text region or the decorative region. This is because the decorative region may contain neighboring connected components of similar thickness.

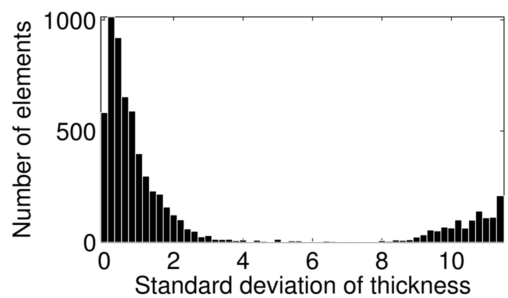


Figure 4. Histogram of the std

By separating the two clusters in this histogram we can discriminate connected components which have high thickness variation with one of their neighbors from the remaining ones. The threshold value T_{std} that separates the two clusters is simply determined by:

$$T_{std} = \frac{\max_{i,j} std(g_i, g_j)}{2} \quad (5)$$

and the adjacent Voronoi regions containing body text components thus should satisfy:

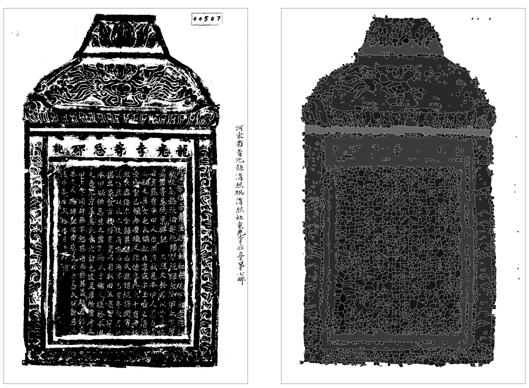
$$std(g_i, g_j) < T_{std} \quad (6)$$

The value of T_{std} is dependent on each image. For the stele image in Figure 1, T_{std} has the value 5.76.

As said before, the standard deviation is sometimes not enough to discriminate all the connected components lying at the decorative region because in that region neighboring connected components may have std less than T_{std} (an example is in Figure 5(b) in which the gray level of each Voronoi region is proportional to the thickness of the connected component belonging to that region). We utilize one more criterion based on the observation that a Voronoi region, which contains a body text component, should not has many adjacent regions as follows:

$$N(g_i) < N_{max} \tag{7}$$

where $N(g_i)$ is the number of adjacent regions of a Voronoi region containing component g_i and N_{max} is a fixed value of 20 determined by experiment.



(a) (b)

Figure 5. Voronoi regions of a stele image

Consequently, in order to extract body text region, we group the Voronoi regions containing text components by utilizing the following procedure:

- *Step 1*: Select the Voronoi region lying at the center of the image, push and insert its label g_i to a queue Q and a list L respectively.
- *Step 2*: Pop a label g_i from Q and select all its adjacent Voronoi regions that have not been selected before. Among these selected Voronoi regions, only keep regions with labels $g_j s$ that satisfy (6) and (7). Push and insert $g_j s$ to Q and L respectively.
- *Step 3*: Iterate *Step 2* until Q gets empty. At this point, L contains the labels of adjacent Voronoi regions containing body text components.

4. Experimental results

The proposed method has been evaluated on a database of 205 stele images of different types of character fonts, sizes and stele decorations provided by EFEO. Figure 6 shows two examples of body text extraction results.

The two input binary stele images are shown in Figure 6(a) and 6(e). The constructed area Voronoi diagrams are shown in Figure 6(b) and 6(f) in which the gray value of each Voronoi region is proportional to the thickness of the connected component belonging to that region. As shown in the figure, the body text regions and decorative regions in Figure 6(b) and 6(f) are composed of Voronoi regions of similar and diverse gray value respectively. Adjacent Voronoi regions spreading from the Voronoi region lying at the center of the image are grouped together using the process described in Section 3.2. Figure 6(c) and 6(g) show the grouped Voronoi regions. The connected components belonging to those grouped regions, which correspond exactly to the body text, are shown in Figure 6(d) and 6(h).

The ground-truth data for each stele image has been defined by hand in which body text region is defined as a set of all Voronoi regions each containing a body text component. The accuracy is defined as the proportion of body text region pixels extracted. The performance of proposed method has an accuracy of 85.81% and is tabulated in Table 1. There are two sources of error: one is over-grouping that is Voronoi regions from decorative region are considered as containing body text components; the other is splitting in which Voronoi regions in body text region are considered as belonging to decorative region.

	EFEO database
Accuracy($\%$)	85.79
Over-grouping($\%$)	11.70
Splitting($\%$)	2.51

Table 1. Performance of body text region extraction from stele images

5. Conclusions

This paper has presented a method for the extraction of body text regions from stele images using area Voronoi diagram and the homogeneity of connected components. Area Voronoi diagram has demonstrated to be effective in representing the neighborhood of connected components in a digital image. Adjacent Voronoi regions are grouped together based on the information about the thickness of neighboring connected components. The experimental results on a number of stele images have shown that the proposed method has achieved high accurate results and is robust to various types of stele.

Future work is to extract each Nom character from the obtained body text regions and represent them in a database for later retrieval. High-level semantic knowledge of these characters will then be utilized to improve the performance of body text region extraction from stele images.

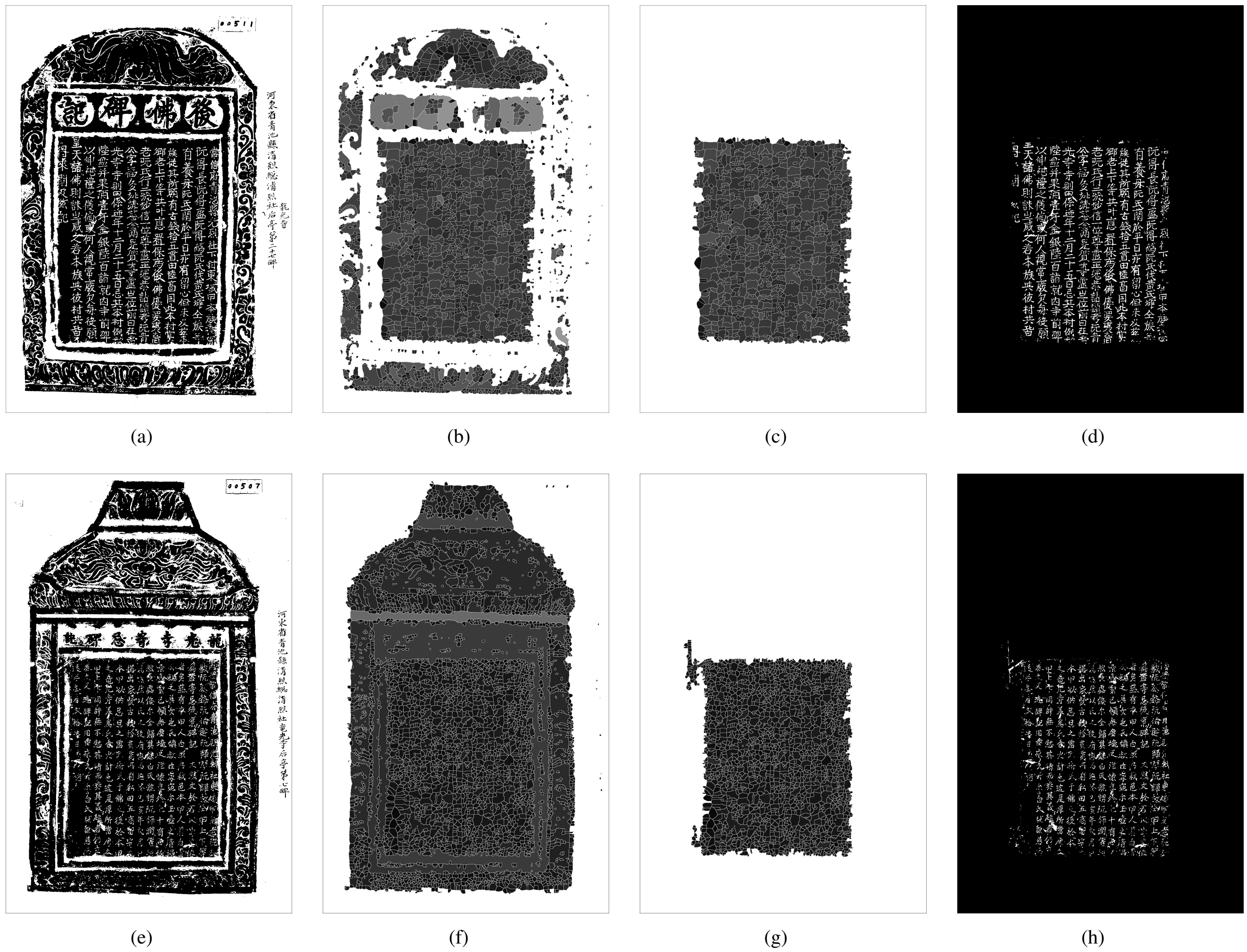


Figure 6. Experimental results of body text extraction from stele images

Acknowledgements

This work is supported by CNRS and EFEO under the framework of an International Scientific Cooperation Program (PICS: 2007-2009).

References

[1] M. Burge and G. Monagan. Using the Voronoi tessellation for grouping words and multi-part symbols in documents. In *Proc. Vision Geometry IV, SPIE*, pages 116–123, 1995.

[2] G. S. di Baja. Well-shaped, stable, and reversible skeletons from the (3,4)-distance transform. *J. Visual Communication and Image Representation*, 5(1):107–115, 2003.

[3] L. A. Fletcher and R. Kasturi. A robust algorithm for text string separation from mixed text/graphics images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):910–918, 1988.

[4] A. K. Jain and S. Bhattacharjee. Text segmentation using Gabor filters for automatic document processing. *Mach. Vision Appl.*, 5(3):169–184, 1992.

[5] K. Kise, A. Sato, and M. Iwata. Segmentation of page images using the area Voronoi diagram. *Comput. Vis. Image Underst.*, 70(3):370–382, 1998.

[6] Y. Lu, Z. Wang, and C. L. Tan. Word grouping in document images based on Voronoi tessellation. In S. Marinai and A. Dengel, editors, *Document Analysis Systems*, volume 3163 of *Lecture Notes in Computer Science*, pages 147–157. Springer, 2004.

[7] Y. Lu, C. Xiao, and C. L. Tan. Constructing area Voronoi diagram based on direct calculation of the Freeman code of expanded contours. *IJPRAI*, 21(5):947–960, 2007.

[8] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial tessellations: Concepts and applications of Voronoi diagrams*. Wiley, NYC, 2000.

[9] Y. Xiao and H. Yan. Text region extraction in a document image based on the Delaunay tessellation. *Pattern Recognition*, 36(3):799–809, 2003.